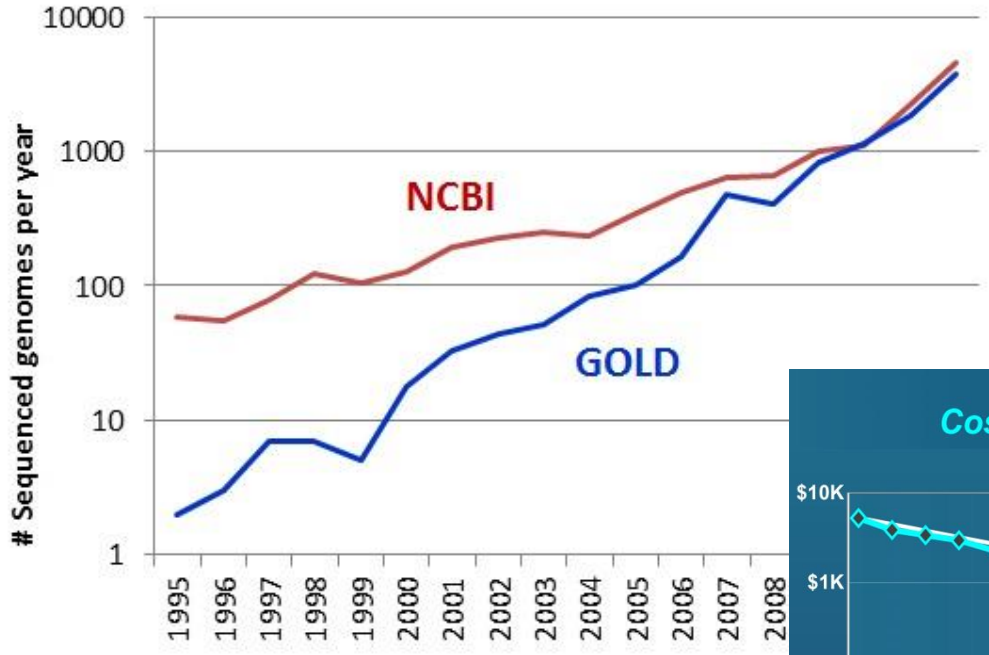
DAY2

Introduction to NGS technologies and library preparation

Chiara Batini, cb334@le.ac.uk

CINECA

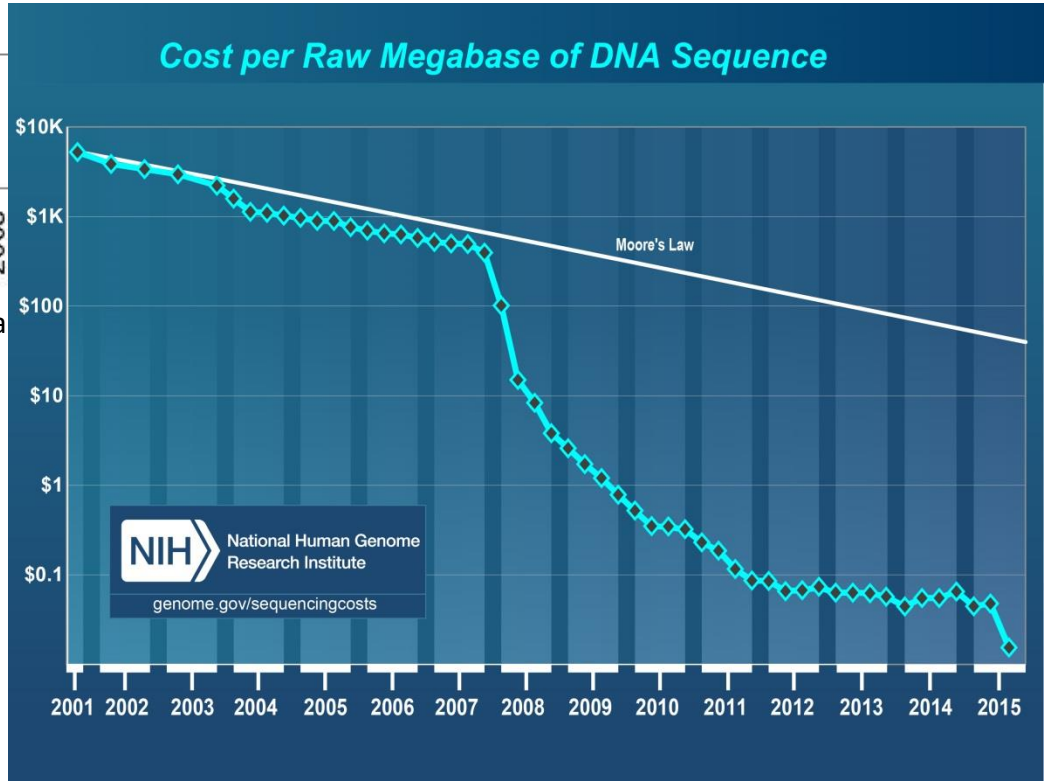
elixir ITALY



from http://sulab.org/2013/06/sequenced-genomes-per-yea

http://www.ncbi.nlm.nih.gov/

https://gold.jgi.doe.gov/



2004: NHGRI launches '$1000 genome' grants

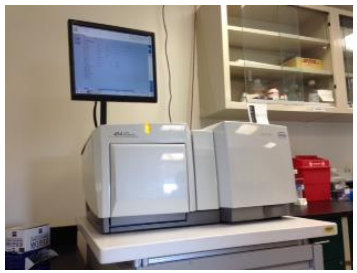
Sequencers: how they were and what they were used for

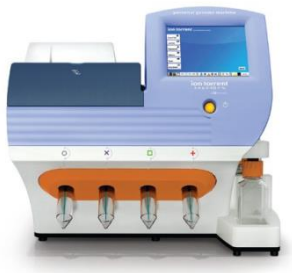
platform	# reads	max read length
Roche 454 – 2005	200K	110bp
Illumina (Solexa) GA – 2006	30M	35bp
ABI SOLiD – 2007	100M	35bp
IonTorrent PGM – 2010	3M	100bp

de-novo assembly
metagenomics

RNA-seq
ChIP-seq
re-sequencing

amplicon-seq



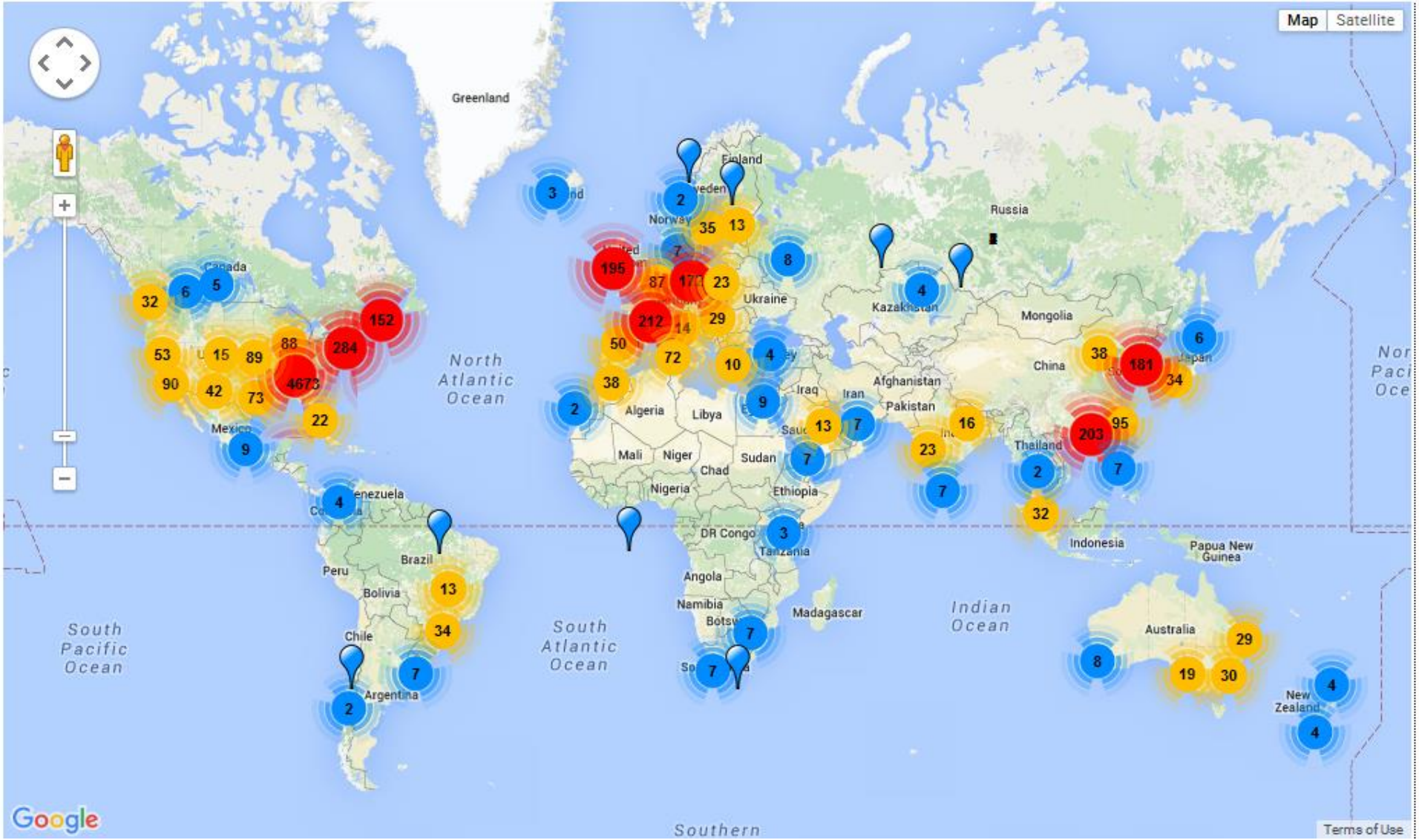




omicsmaps.com/#

Search

Next Generation Genomics: World Map of High-throughput Sequencers

Show all platforms | 454 | HiSeq | HiSeq X Ten | Illumina GA2 | Ion Torrent | MiSeq | MinION | NextSeq | PacBio | Polonator | Proton | SOLiD | Serv

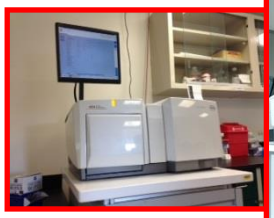


Sequencers: size and output

unit

bench

Roche 454 GS



Roche 454 GS FLX+

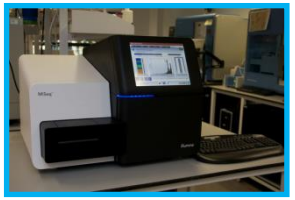
Illumina GA

Illumina HiSeq

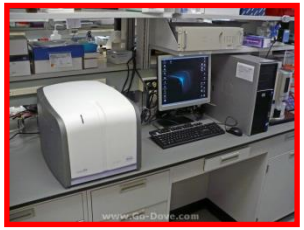
Illumina HiSeq X

Illumina NextSeq

Illumina MiSeq



Roche 454 GS Junior



Ion PGM

Ion Proton

Illumina offers:
highest throughput per run
lowest per-base cost

IonTorrent is the fastest

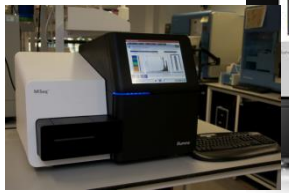
Mb Gb Tb

Sequencers: read length

Ion PGM

Ion Proton

Illumina MiSeq

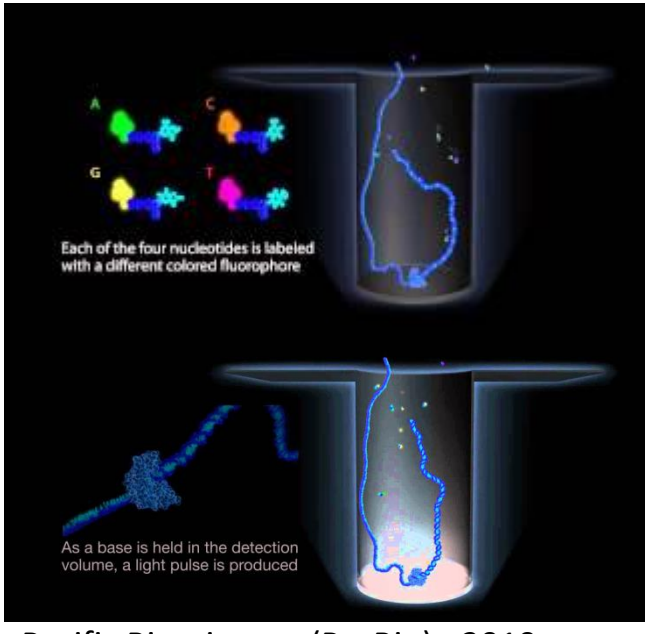


Illumina HiSeq

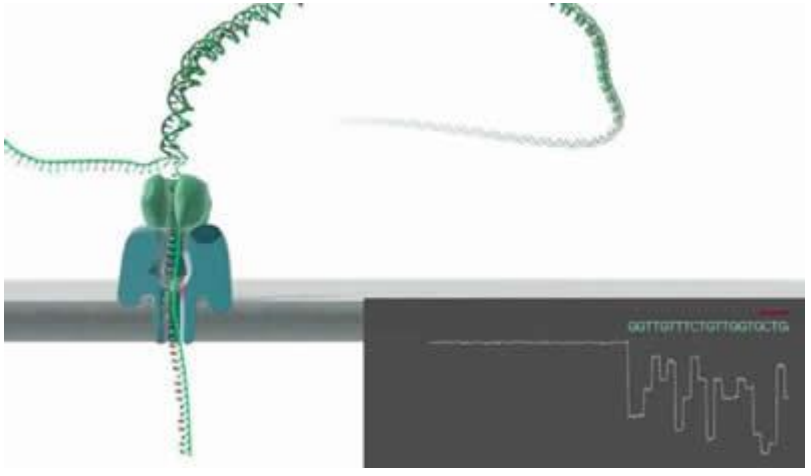
Roche 454 GS FLX+

250-300bp | 400bp | 1kb

Sequencers: the real-time generation



Pacific Biosciences (PacBio) - 2010



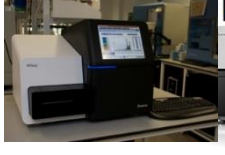
Nanopore MinION - 2014

Ion PGM

Ion Proton

PacBio

Illumina MiSeq



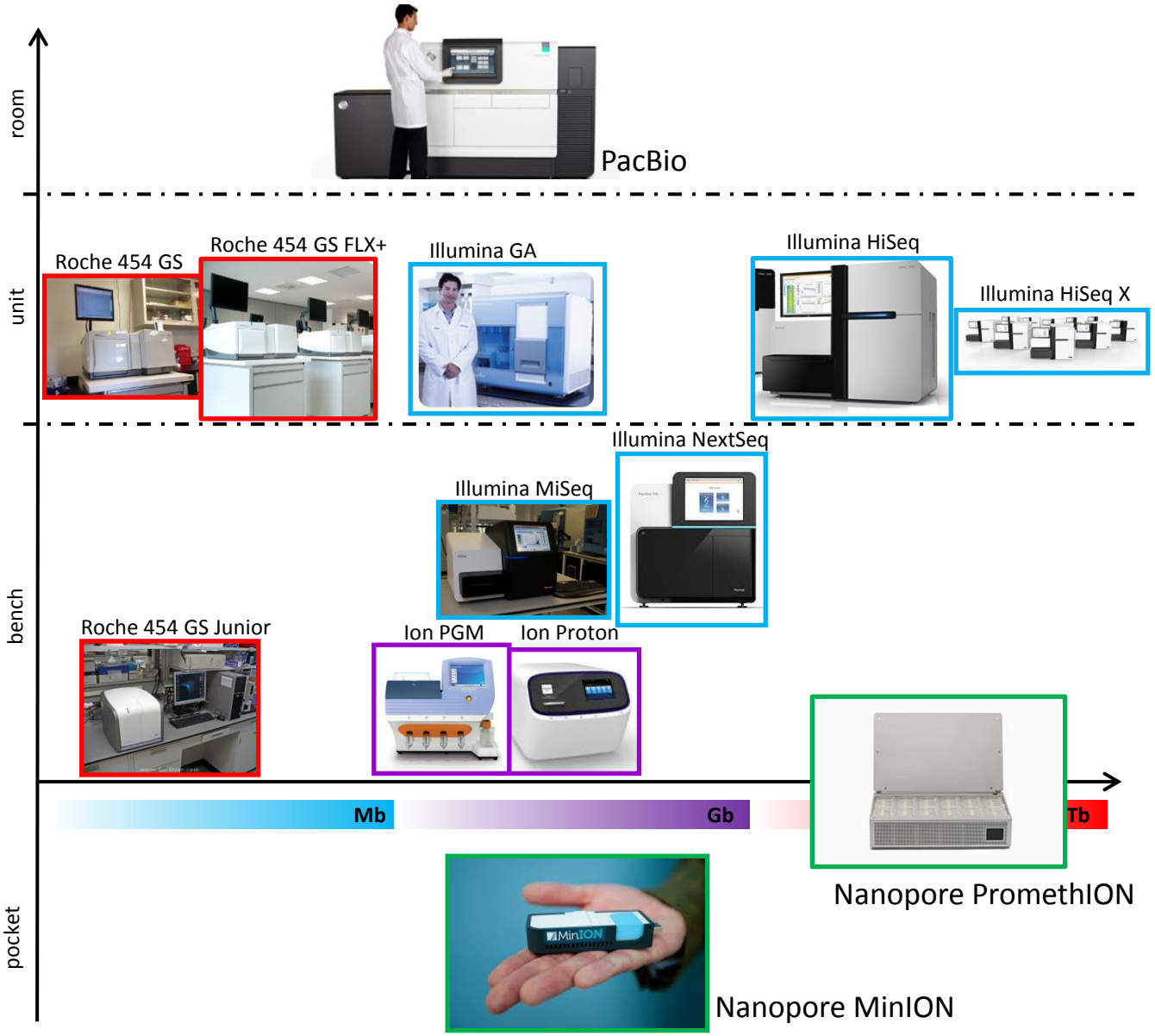
Illumina HiSeq

Roche 454 GS FLX+

Nanopore MinION

250-300bp | 400bp | 1kb | >10kb

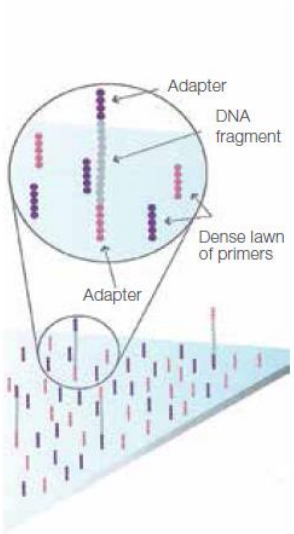
Sequencers: size and output



Illumina sequencing technology (1)

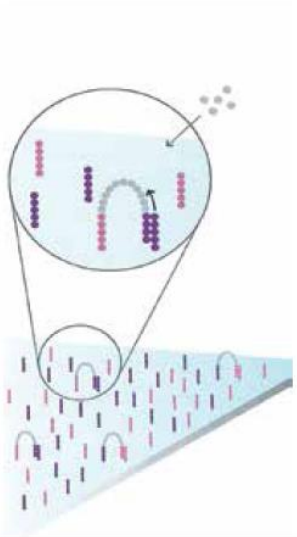
Sequencing by synthesis, 4 modified dNTPs together, imaging support: slide

Figure 3: Attach DNA to Surface



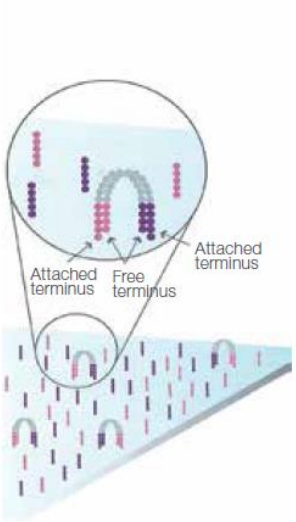
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Figure 4: Bridge Amplification



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Figure 5: Fragments Become Double Stranded

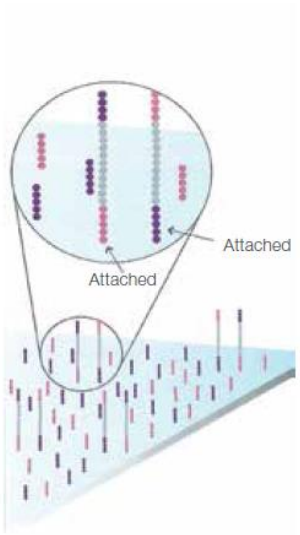


The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

from https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

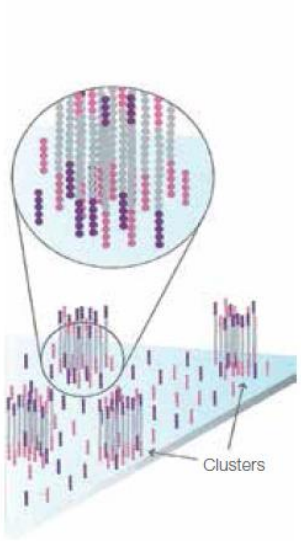
Illumina sequencing technology (2)

Figure 6: Denature the Double-Standed Molecules



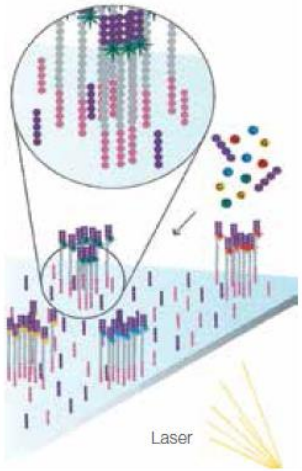
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 8: Determine First Base



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

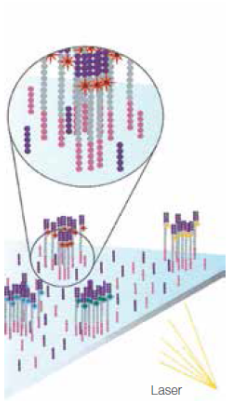
from https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina sequencing technology (3)

Figure 9: Image First Base

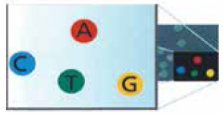
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Figure 10: Determine Second Base



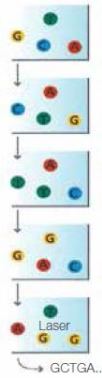
The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 11: Image Second Chemistry Cycle



After laser excitation, the image is captured as before, and the identity of the second base is recorded.

Figure 12: Sequencing Over Multiple Chemistry Cycles

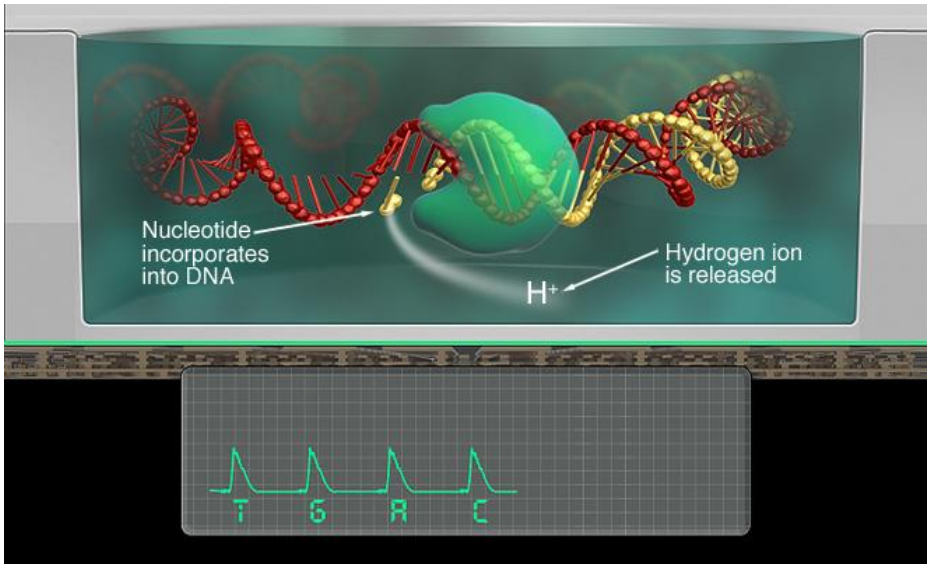


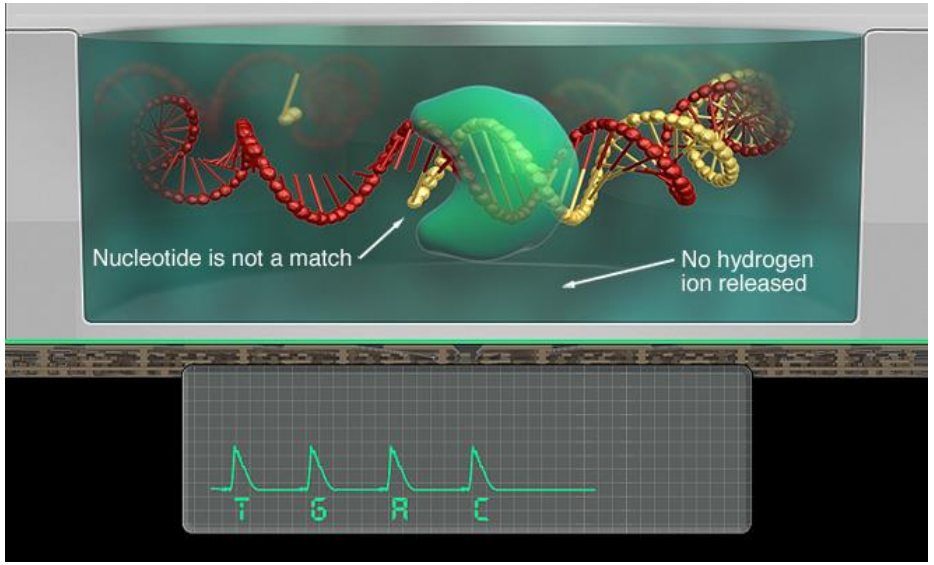
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

from https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

IonTorrent sequencing technology (1)

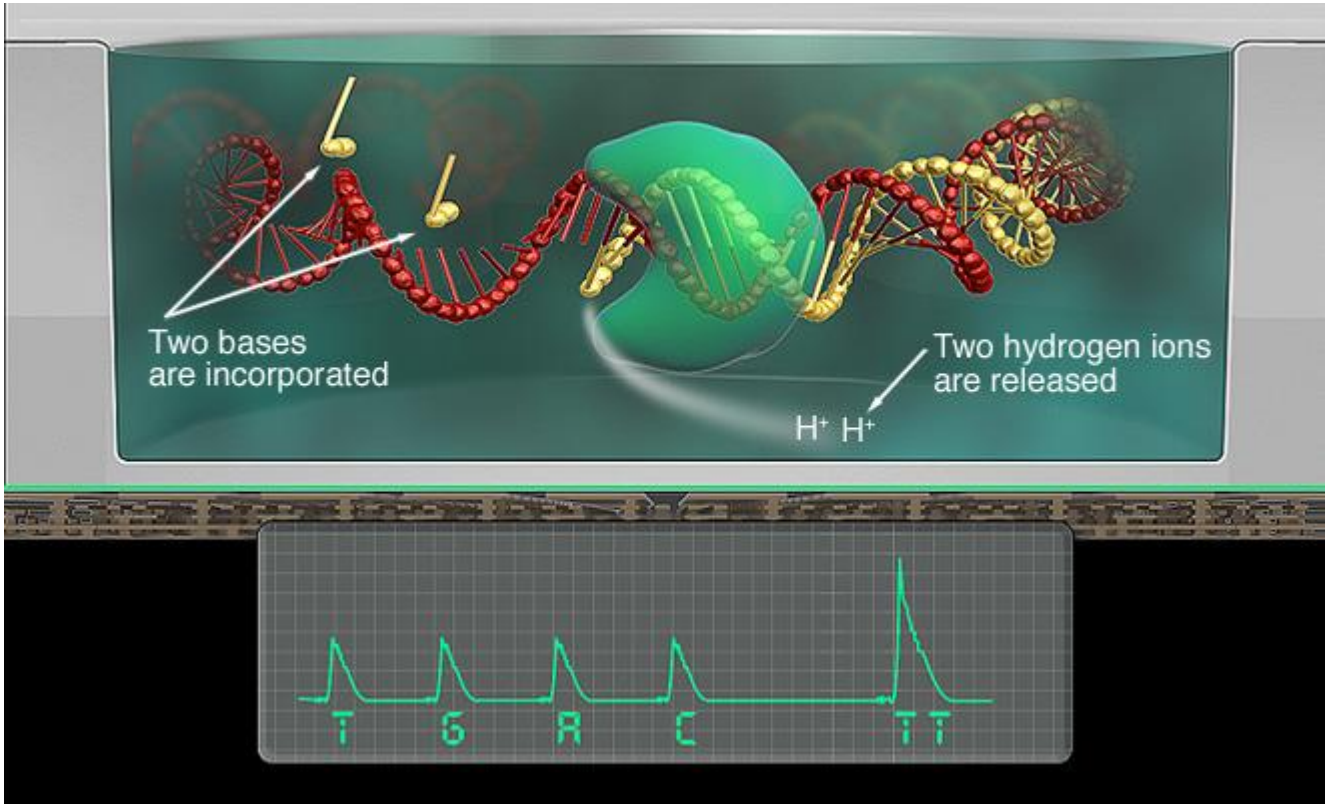
Sequencing by synthesis, 1 non-modified dNTP at the time, pH detection
support: bead/chip





from https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

IonTorrent sequencing technology (2)



from https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

Sequencers: error rates and patterns

Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
ABI 3730xl (capillary)	substitution	0.1-1	0.1-1
Roche 454 – All models	indel	1	1
Illumina – All models	substitution	~0.1	~0.1
Ion Torrent – all chips	Indel	~1	~1
Oxford Nanopore	deletions	≥4*	4*
PacBio RS	Indel	~13	≤1

from http://www.molecularecologist.com/next-gen-fieldguide-2014/

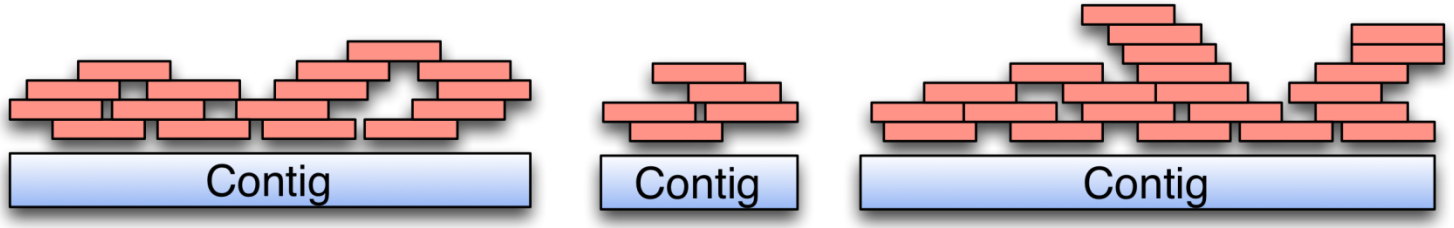
NGS applications

Genomic DNA:

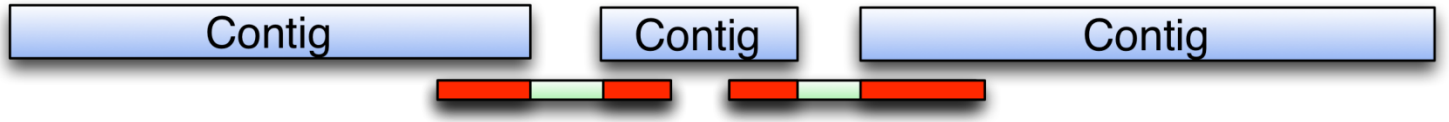
- WGS (whole genome sequencing): population-scale; single-cell
- WES (whole exome sequencing)
- custom enrichment
- amplicon-seq
- RAD-seq (restriction-site-associated DNA): very useful for organisms lacking reference genomes

De novo assembly

Overlap reads by sequence, into a single contiguous sequence (Contig)

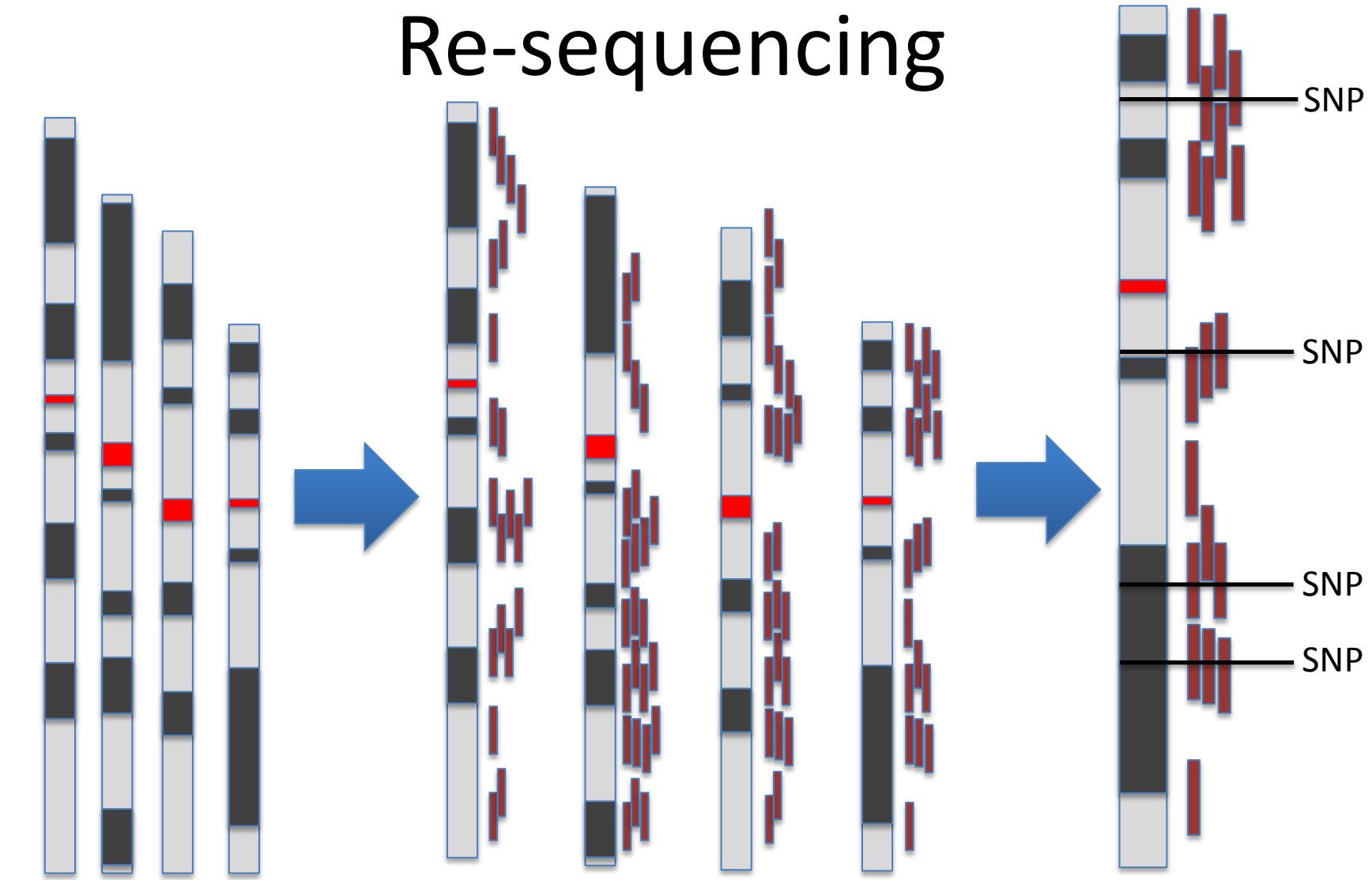


Fill gaps using known contig end sequences



Finished chromosome sequence

Re-sequencing



What does DNA-seq tell you

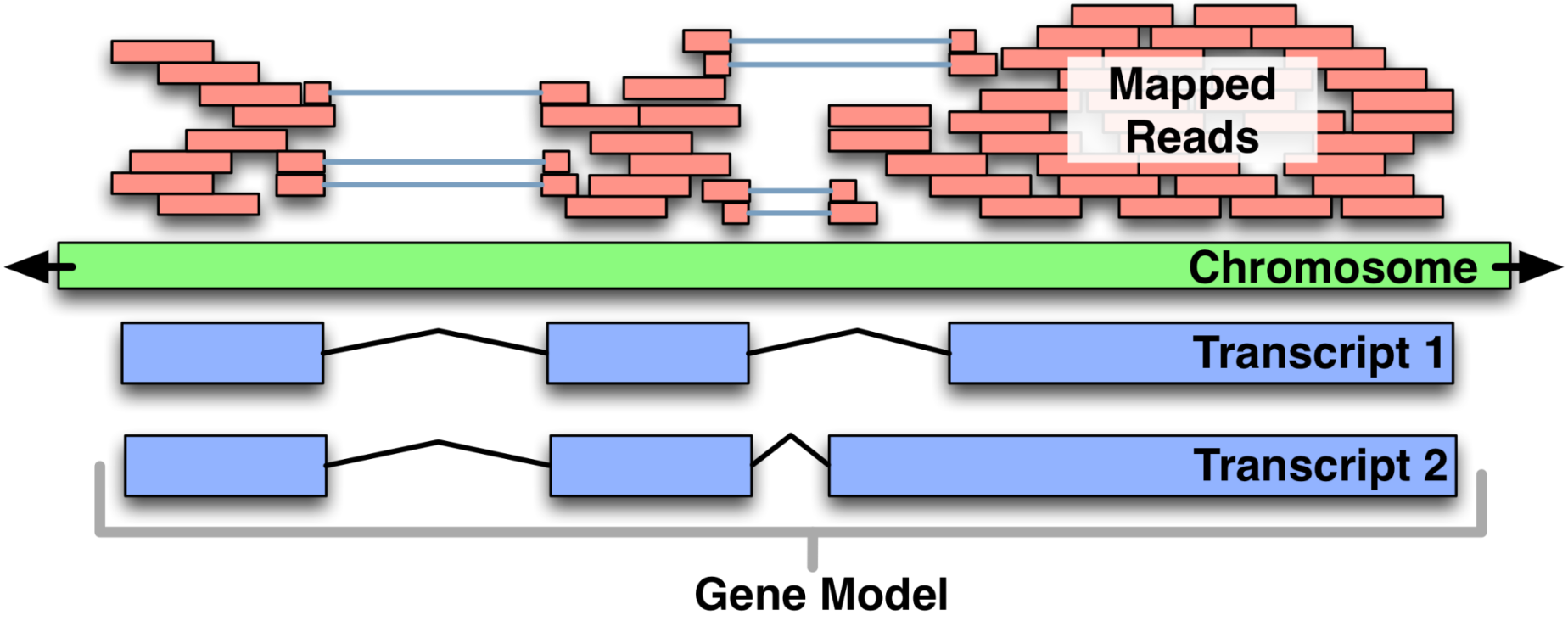
- Calling of known polymorphisms
- Identification of novel polymorphism (SNPs, indels etc)
- Genomic rearrangements, large deletions and insertions
- What the genome of your sample looks like (genome assembly, chromosome assembly)

NGS applications

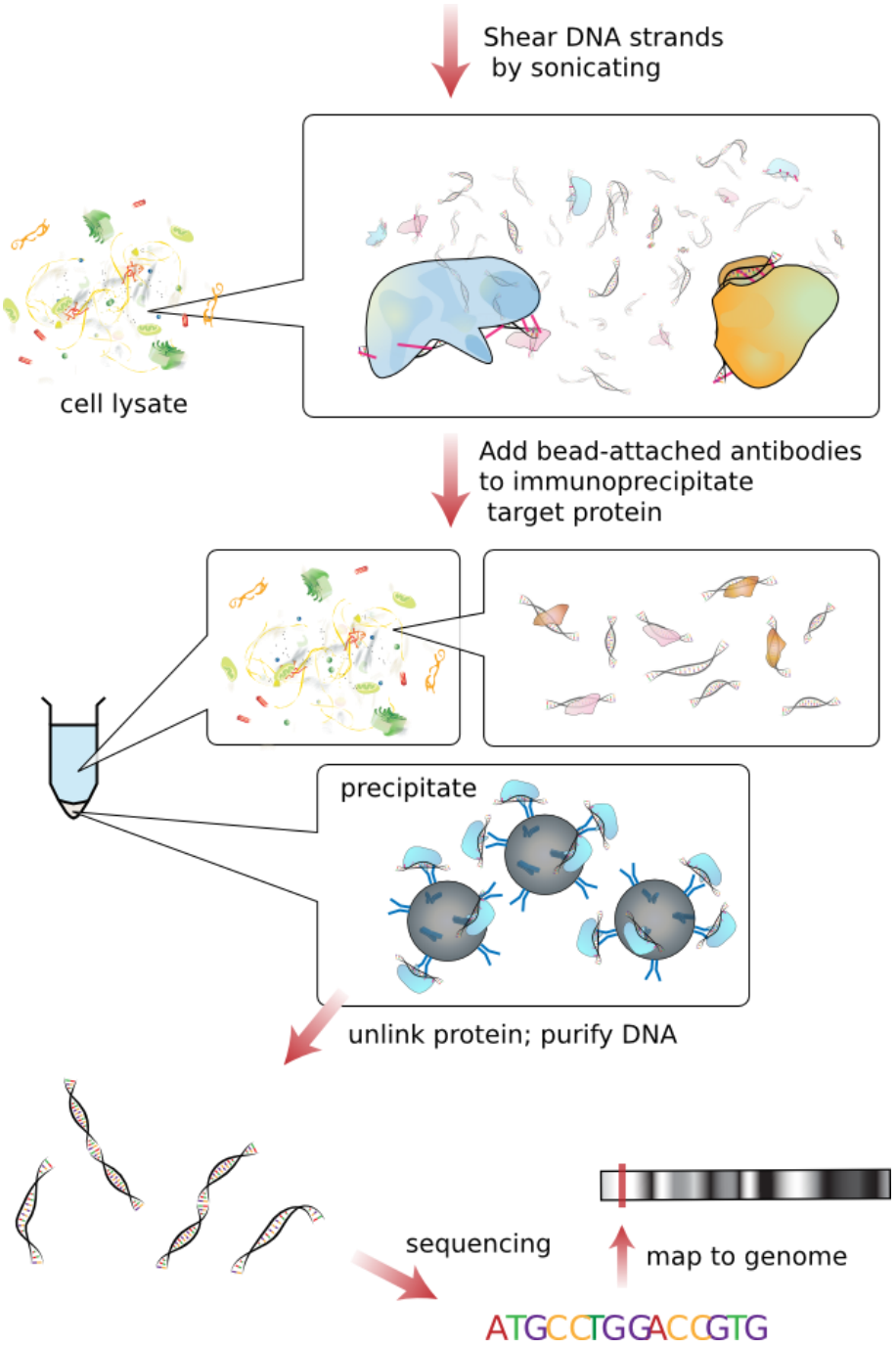
RNA-seq:

- strand-specific RNA-seq protocols
- single-cell transcriptomics
- Fluorescent in situ RNA seq (FISSEQ): it allows the characterization of the single-cell transcriptomics and the localization of the transcripts withi the cell
- CaptureSeq
- Native elongating transcript seq (NET-seq)

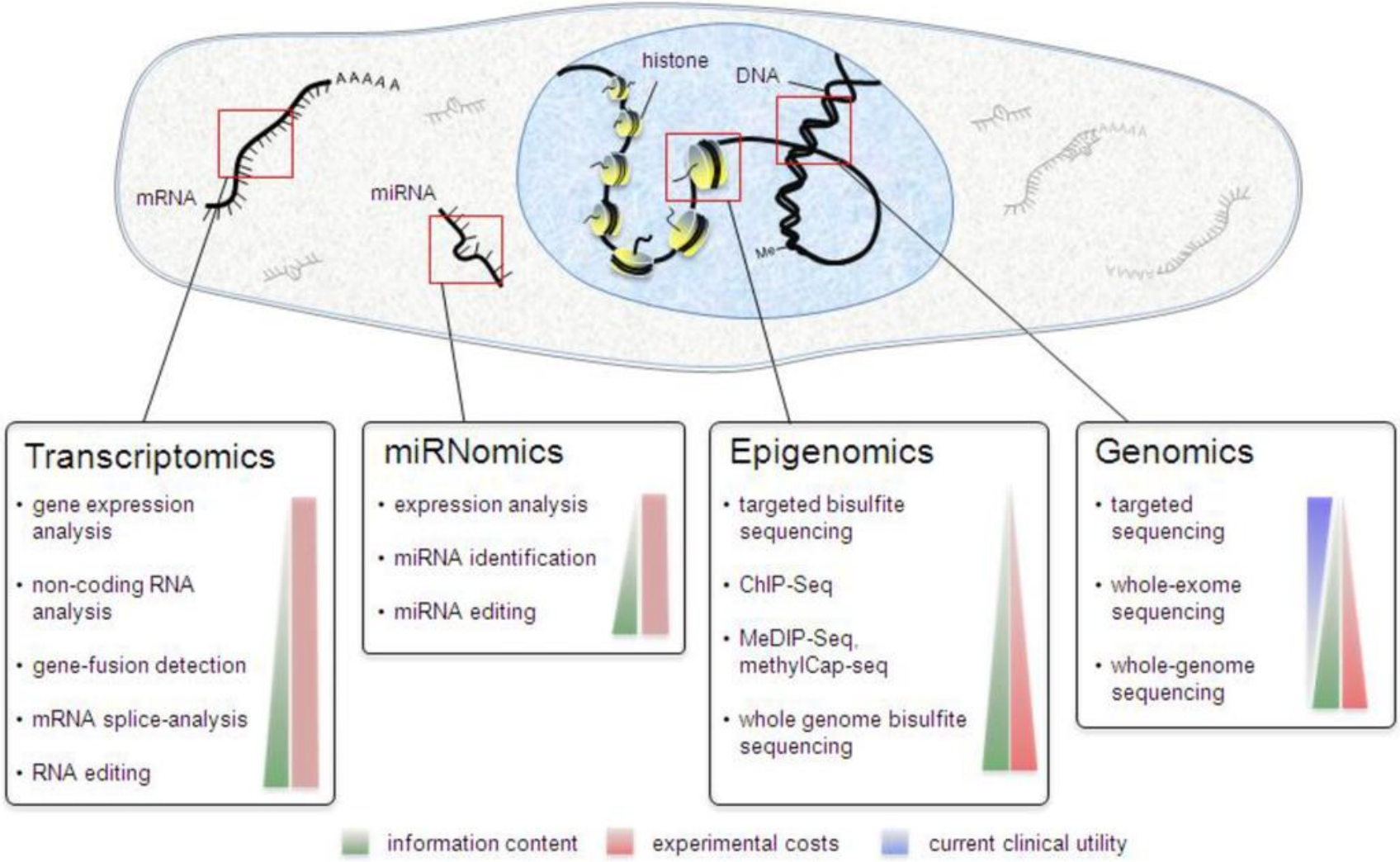
RNA-seq


Mapped
Reads
Chromosome
Transcript 1
Transcript 2
Gene Model

ChIP-seq



Location-based techniques
ChIP-seq: protein-DNA interactions; ChIP-exo (down to nucleotide level)
protein-RNA, RNA-DNA, DNA-DNA interactions



from http://www.mdpi.com/2079-7737/2/1/378/htm

Diagnostic tools in clinical laboratories; spreading in forensics

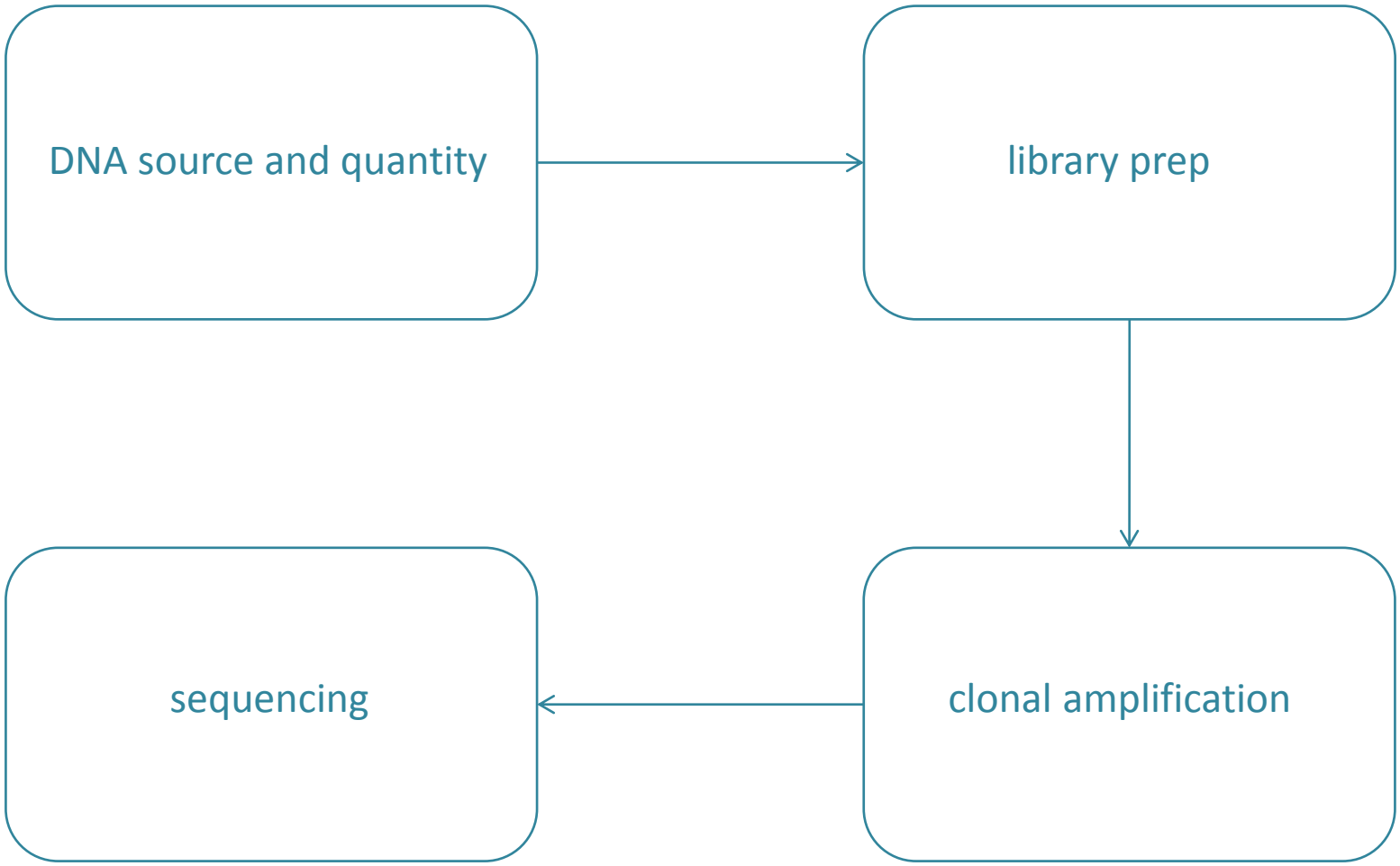
Illumina HiSeq X Ten : a human genome for less than $1000?
Personalized medicine is just around the corner…

So, which is the best technology?

https://docs.google.com/spreadsheets/d/1eoWVtwKvnbV8MaRb9hzCSpGtEjazO0XTt3a8TbPG4dY/edit?pli=1#gid=0

Good support and communication are essential

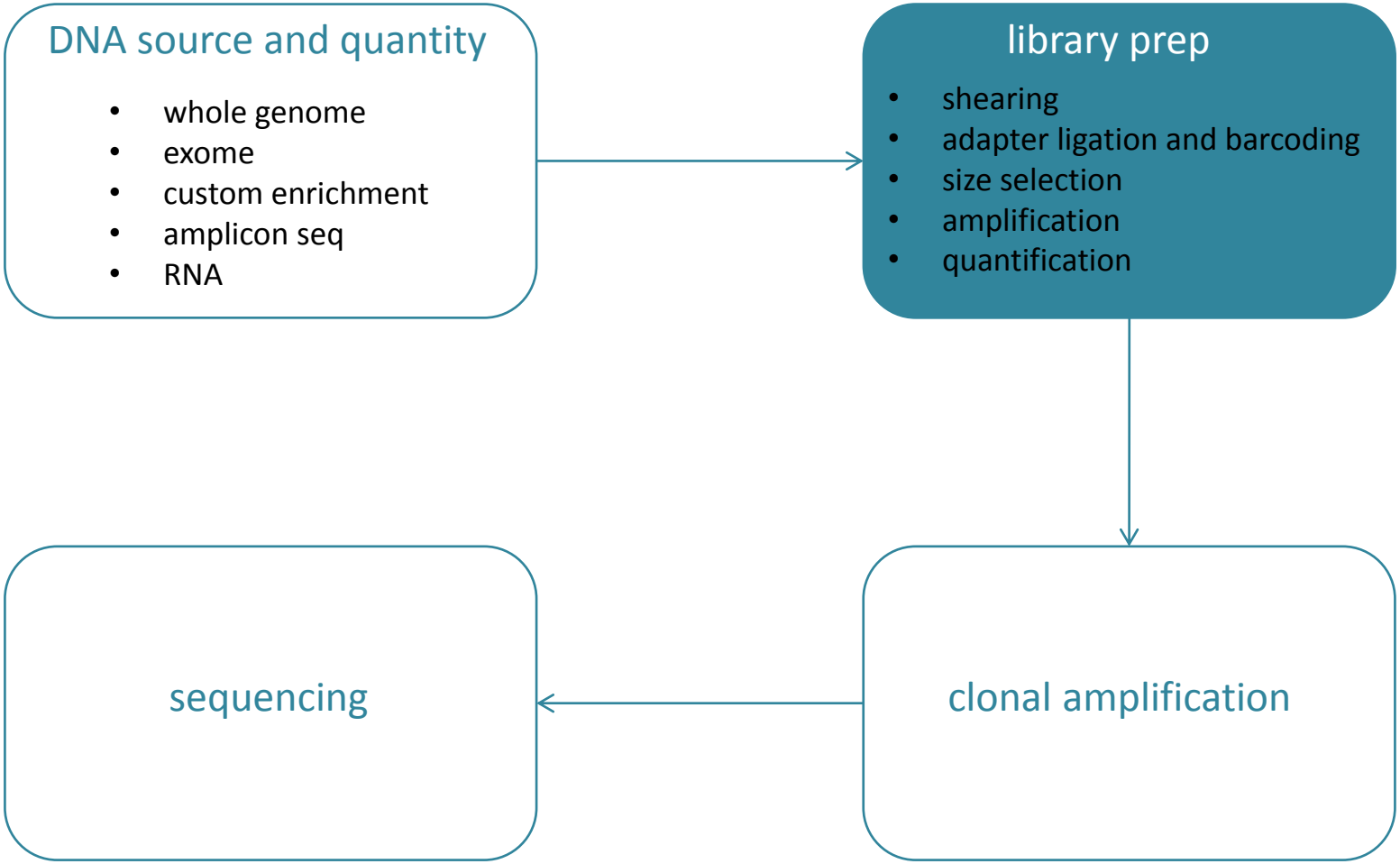
So, how do we get our DNA ready to be sequenced then?


DNA source and quantity
library prep
clonal amplification
sequencing

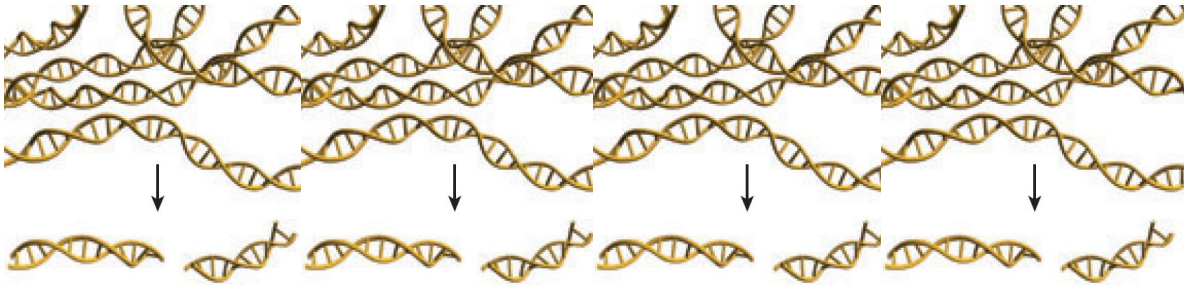
DNA source and quantity

non-degraded, RNA-free, high quality DNA
quality can be checked with nanodrop, specific ratios are suggested:
260/280: 1.7-2.0
260/230: >2
for quantification other methods are suggested [Qubit, picogreen]

whole genome	2ug for paired end, 20ug for mate pair
Exome	6ug or more
custom enrichment	min 5ug, more depending on the target size
amplicon seq	a few ng per amplicon may be enough, since PCR will increase the concentration


DNA source and quantity
whole genome
exome
custom enrichment
amplicon seq
RNA
library prep
shearing
adapter ligation and barcoding
size selection
amplification
quantification
clonal amplification
sequencing

library prep - shearing



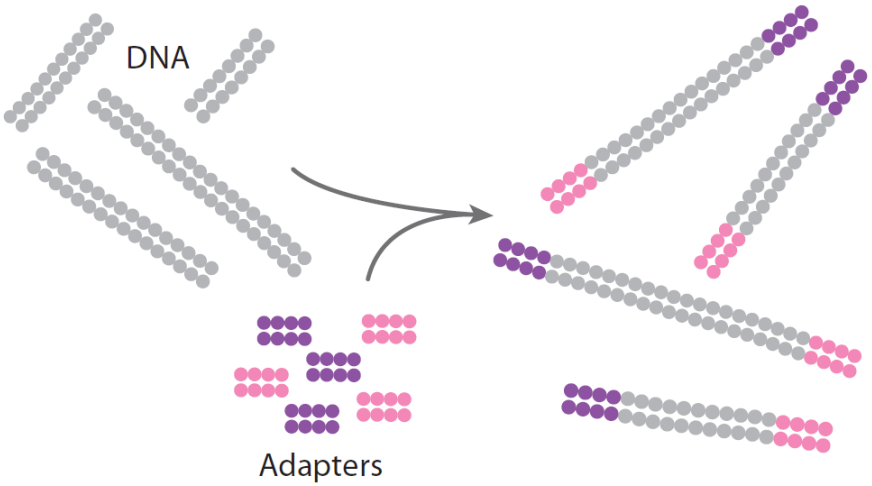
modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

Two main methods are used:

- Sonication: hydrodynamic shearing using acoustic energy; bubbles are formed in solution, when they explode they break the DNA

- Enzymatic reaction: enzymes randomly cut the dsDNA in a time-dependent manner

library prep – adapter ligation and barcoding

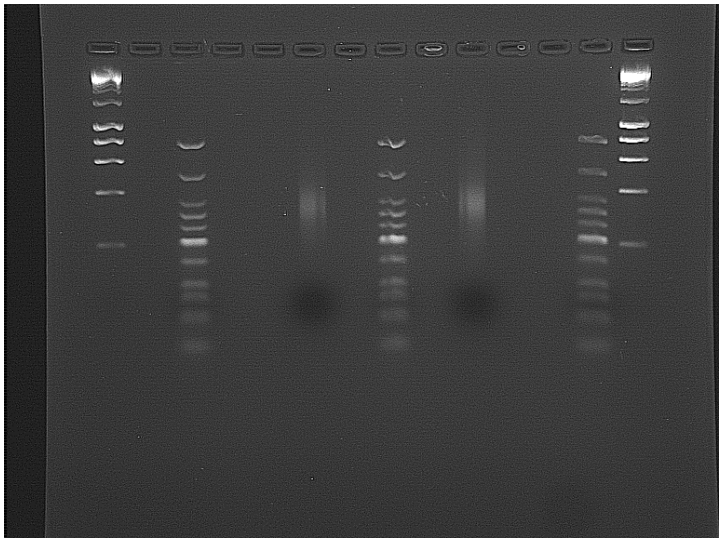


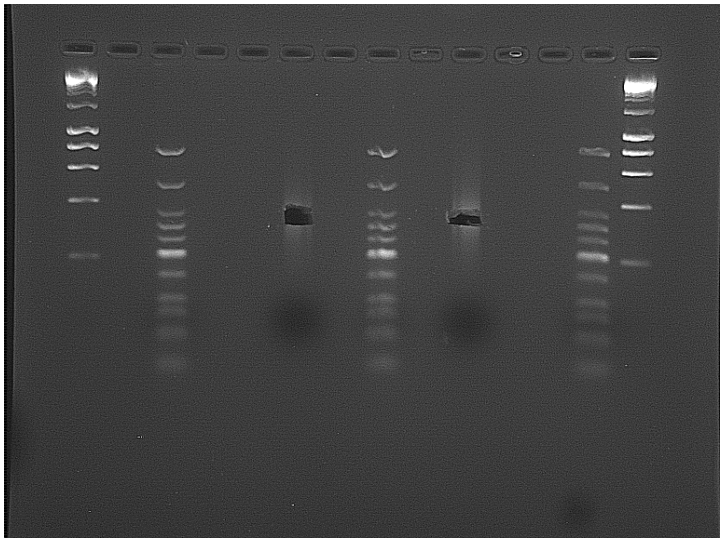
modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

- adapters: 30-50bp fragments which contain primer sites for amplification and are needed to link the fragment with the support (slide, bead)
- barcodes/indexes: 6-10bp fragments which carry a unique sequence; they are used to distinguish samples run in the same lane/chip
- Adapters and barcodes are combined into one fragment

library prep – size selection

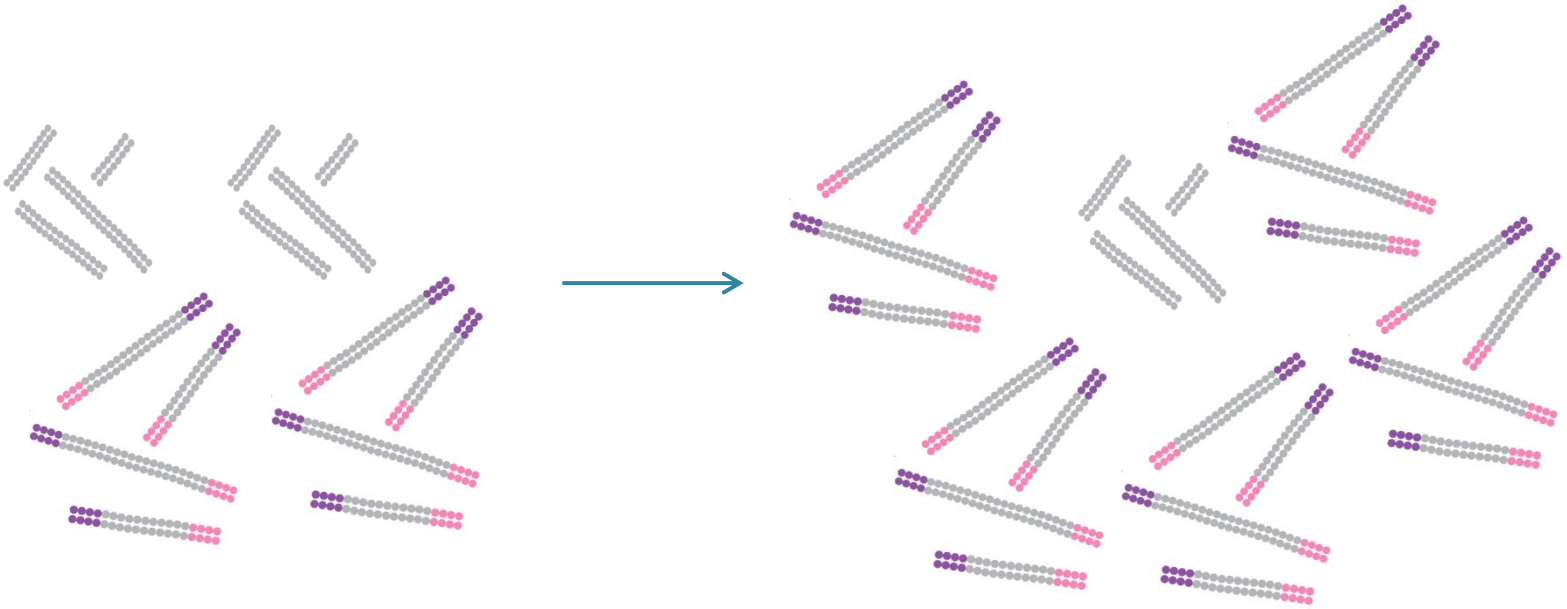
- Needed to select fragments of the right length from a mixture of fragments generated by the fragmentation step: it is usually 200-300bp for paired end sequencing with Illumina, but it may vary depending on the read length for 454 and IonTorrent





library prep – amplification

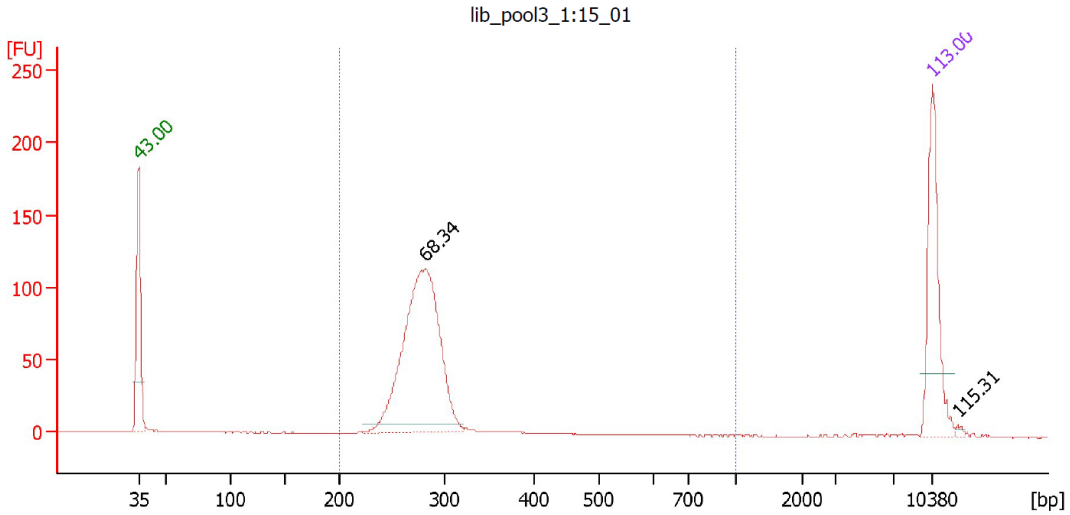
- Amplification: needed to increase the concentration of the fragments which positively incorporated adapters

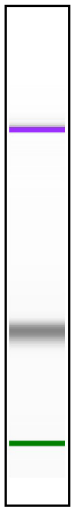


modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

library prep –quantification

- Quantification: needed to tune the quantity of template for the run – usually performed with the Agilent Bioanalyzer



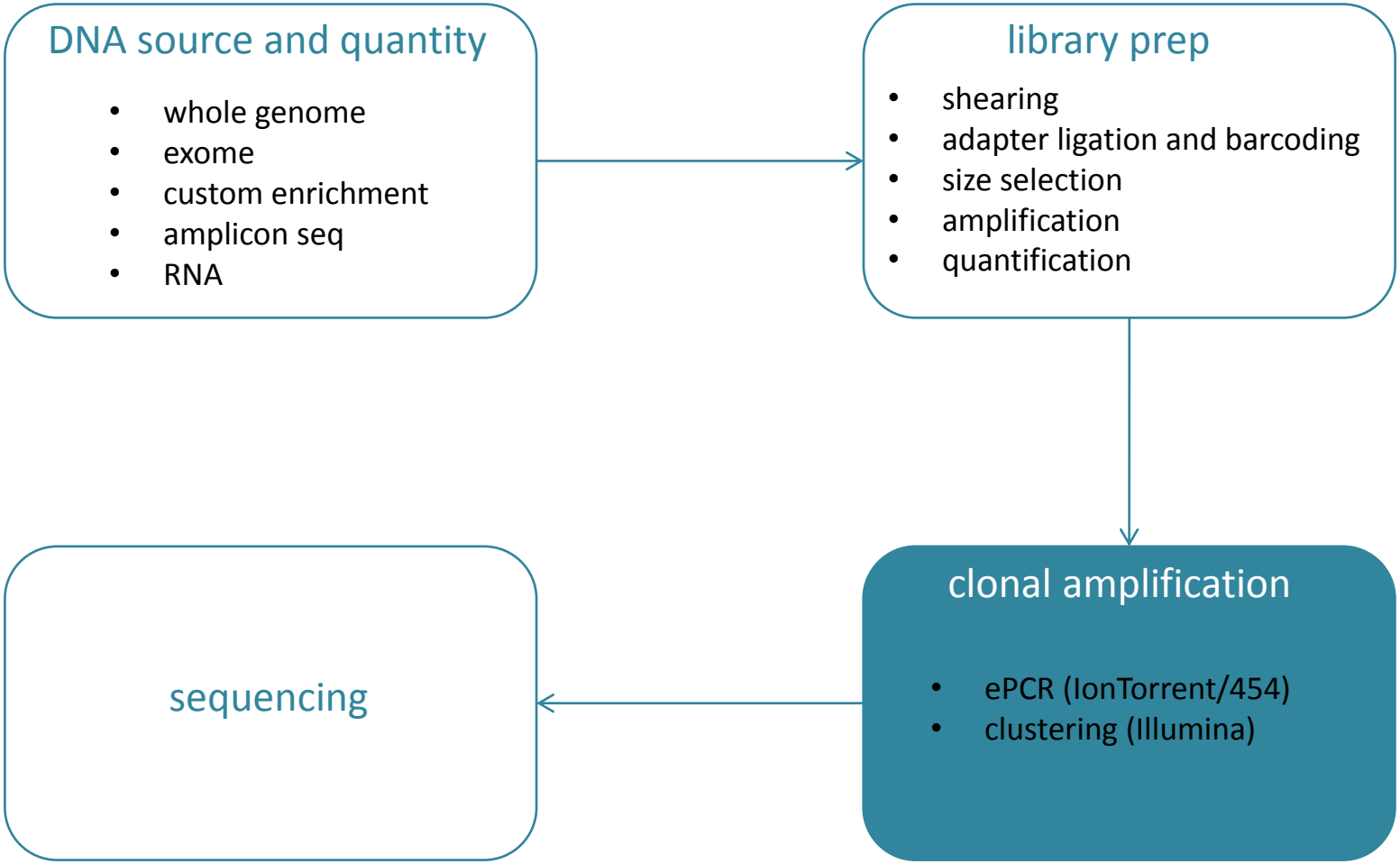


Overall Results for sample 3 : lib_pool3_1:15_01

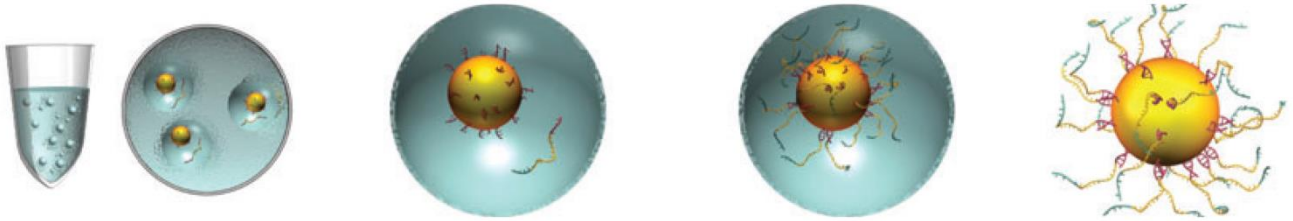
Number of peaks found:	2	Corr. Area 1:	695.4
Noise:	0.2		

Peak table for sample 3 : lib_pool3_1:15_01

Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations	Area
1	35	125.00	5,411.2	Lower Marker	83.8
2	282	335.56	1,802.2		460.4
3	10,380	75.00	10.9	Upper Marker	278.6
4	12,706	0.00	0.0		6.2


DNA source and quantity
whole genome
exome
custom enrichment
amplicon seq
RNA
library prep
shearing
adapter ligation and barcoding
size selection
amplification
quantification
clonal amplification
ePCR (IonTorrent/454)
clustering (Illumina)
sequencing

clonal amplification – ePCR (Roche 454/Ion Torrent)



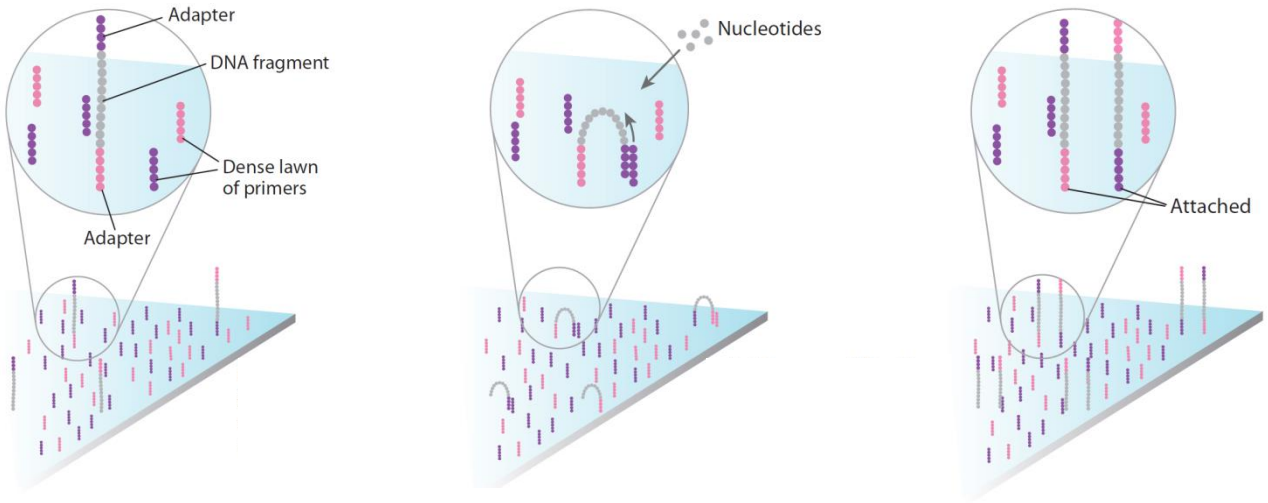
modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

emulsion PCR: ideally each drop contains a single fragment, PCR reagents and a bead with primers [beads may happen to be polyclonal, will be discarded after sequencing]

each bead will have one million copies of each fragment on its surface

an enrichment step is needed afterwards in order to increase the concentration of beads with attached fragments

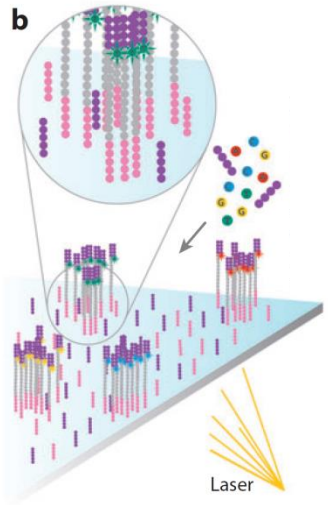
clonal amplification – clustering Illumina



modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

isothermal amplification: temperature is kept constant but reagent are cycled to perform amplification, not very efficient (on purpose)

clonal amplification – clustering Illumina



modified from Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

clusters usually contain around 1000 identical copies of a single template

optimal cluster density will provide optimal sequencing output

A few videos…

Illumina
https://www.youtube.com/watch?v=womKfikWlxM

IonTorrent
https://www.youtube.com/watch?v=ZL7DXFPz8rU

PacBio
https://www.youtube.com/watch?v=NHCJ8PtYCFc

Oxford Nanopore
https://www.youtube.com/watch?v=3UHw22hBpAk

useful references and links

Mardis 2008 Annu Rev Genomics Hum Genet 9:387-402

Metzker 2010 Nat Rev Genet 11:31-46

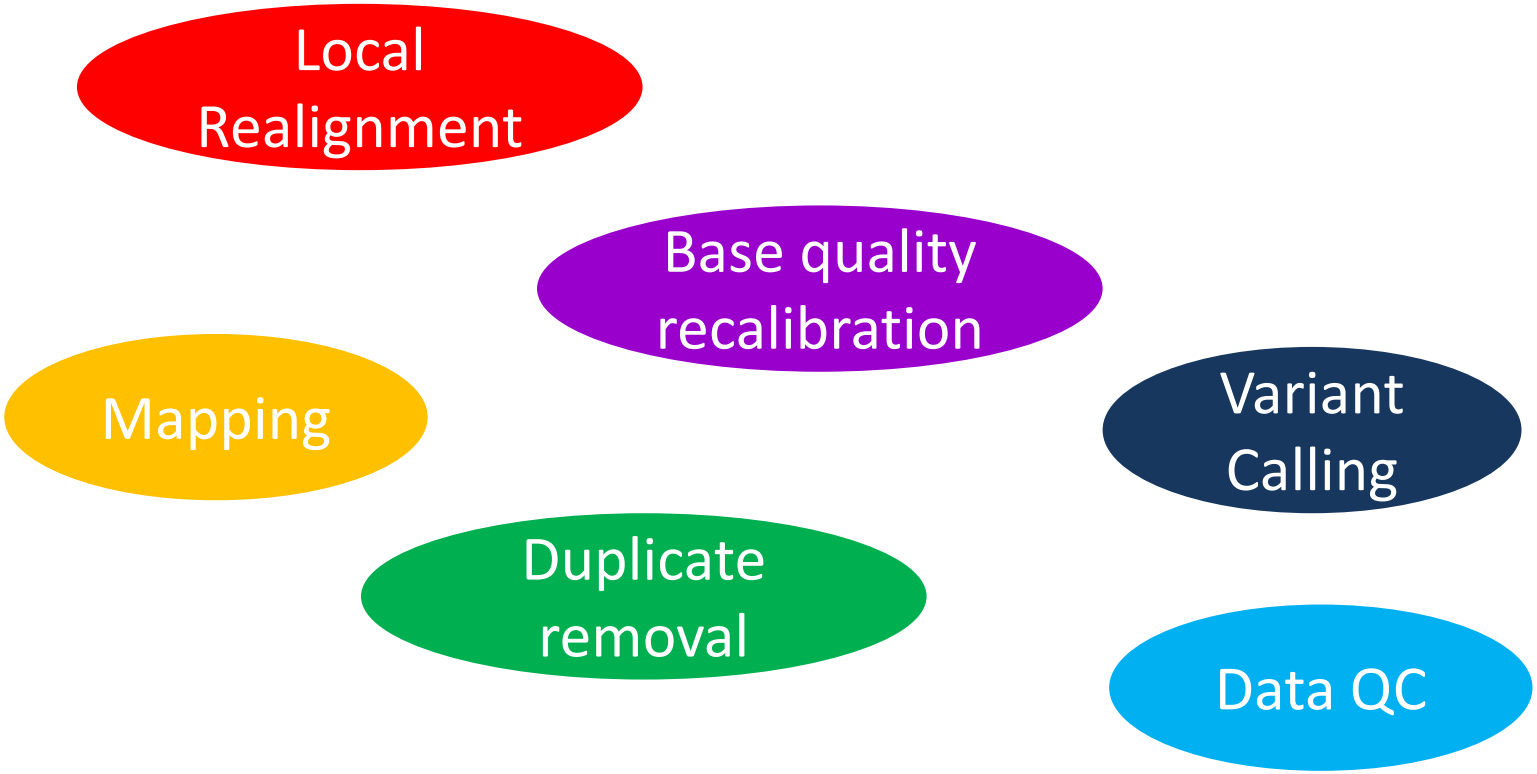
http://seqanswers.com/forums/index.php

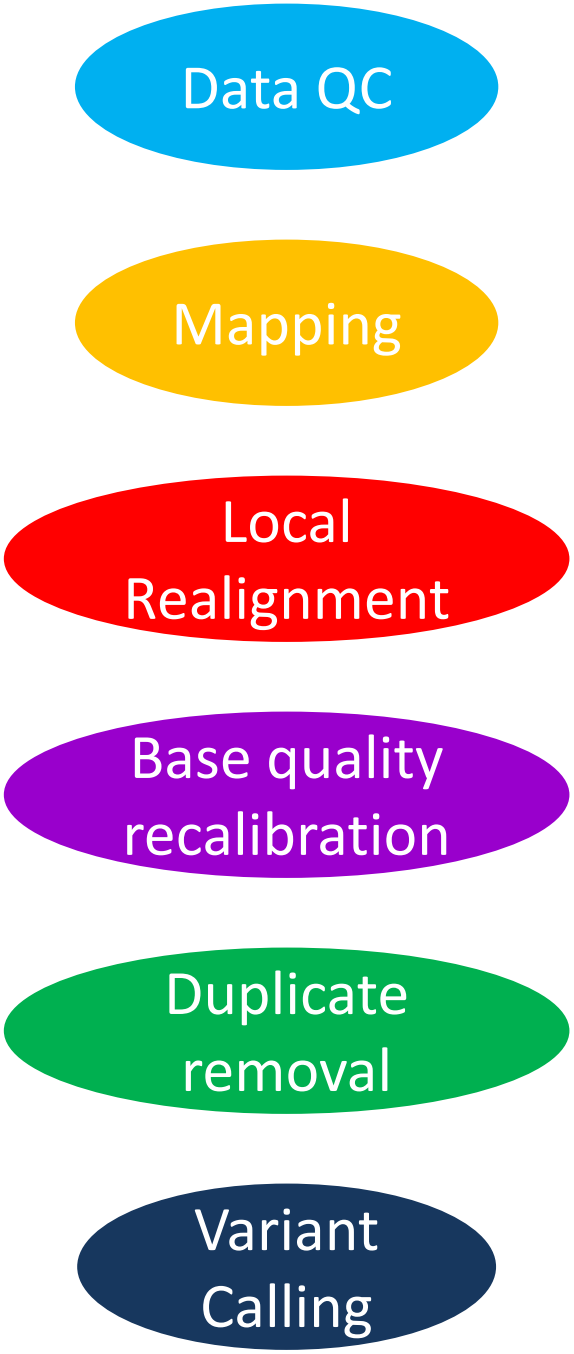
http://www.frontlinegenomics.com/1649/next-generation-sequencing-how-and-why-we-got-here/

http://thewestheimerinstitute.org/pubs/The%20challenges%20of%20sequencing%20by%20synthesis.pdf

http://www.molecularecologist.com/next-gen-fieldguide-2014/

Thanks to:
Pierpaolo Maisano Delser
Rita Neumann
James Eales


Local Realignment
Base quality recalibration
Mapping
Variant Calling
Duplicate removal
Data QC


Data QC
Mapping
Local Realignment
Base quality recalibration
Duplicate removal
Variant Calling